1770 · 65128
(312) 360 0080

日 本 国 特 許 庁
PATENT OFFICE
JAPANESE GOVERNMENT

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日
Date of Application: ２０００年 ５月２６日

出 願 番 号
Application Number: 特願２０００－１５５８６７

出 願 人
Applicant (s): 富士通株式会社

CERTIFIED COPY OF
PRIORITY DOCUMENT

２０００年１０月２７日

特 許 庁 長 官
Commissioner,
Patent Office 及 川 耕 造

出証番号 出証特２０００－３０８９２７７

【書類名】　特許願

【整理番号】　0050852

【提出日】　平成12年 5月26日

【あて先】　特許庁長官殿

【国際特許分類】　G06F 17/30

【発明の名称】　文書情報検索装置、方法及び文書情報検索プログラムを格納したコンピュータ可読の記録媒体

【請求項の数】　10

【発明者】

【住所又は居所】　神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内

【氏名】　阿部　静一郎

【特許出願人】

【識別番号】　000005223

【氏名又は名称】　富士通株式会社

【代理人】

【識別番号】　100079359

【住所又は居所】　東京都港区西新橋３丁目２５番４７号　清水ビル８階

【弁理士】

【氏名又は名称】　竹内　進

【電話番号】　03(3432)1007

【選任した代理人】

【識別番号】　100093584

【住所又は居所】　東京都港区西新橋３丁目２５番４７号　清水ビル８階

【弁理士】

【氏名又は名称】　宮内　佐一郎

【電話番号】　03(3432)1007

【手数料の表示】

【予納台帳番号】　009287

【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【包括委任状番号】　9704823

【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　文書情報検索装置、方法及び文書情報検索プログラムを格納したコンピュータ可読の記録媒体

【特許請求の範囲】

【請求項１】

ネットワークを経由した検索要求に基づいて文書情報を検索して応答する文書情報検索装置に於いて、

前記検索要求元に、検索条件としてファイルを指定し、指定したファイル内容をネットワークを経由して送信する検索条件指定部を設け、

検索側に、前記検索条件指定部から送信されたファイル内容からキーワードを生成してデータベースから類似文書を検索する検索マシンを設けたことを特徴とする文書情報検索装置。

【請求項２】

請求項１記載の文書情報検索装置に於いて、

前記データベースは、検索対象文書から抽出した重要語を列挙したインデックス情報を文書毎に保存し、

前記検索マシンは、

検索要求に伴って受信したファイル内容からテキスト文を抽出するテキスト抽出処理部と、

前記テキスト文の形態素解析により名詞を抽出する形態素回析部と、

前記名詞の中から重要語を抽出して論理和でつなげたキーワードを生成するキーワード生成部と、

前記キーワードによる検索データベースの検索で類似する文書を検索してクライアントに検索結果を通知する検索実行部と、

を備えたことを特徴とする文書情報検索装置。

【請求項３】

請求項２記載の文書情報検索装置に於いて、前記キーワード生成部は、各名詞が前記文書データベースに格納した検索文書毎のインデックス中の何文書に出現するかの出現数をカウントし、所定範囲の出現数をもつ上位の所定数の単語を選択してキーワードを生成することを特徴とする文書情報検索装置。

【請求項４】

請求項３記載の文書情報検索装置に於いて、前記キーワード生成部は、インデックス中の文書数Ｎとした場合、出現数Ｈが

$$2N/3 \geqq H \geqq 1$$

の範囲の出現数をもつ上位の１０個の単語を選択してキーワードを生成することを特徴とする文書情報検索装置。

【請求項５】

請求項１記載の文書情報検索装置に於いて、前記キーワード生成部は検索要求に伴って受信したファイルから抽出したプロパティ情報を前記キーワードに含めて検索させることを特徴とする文書情報検索装置。

【請求項６】

検索対象文書から抽出した重要語を列挙したインデックス情報を文書毎に保存しているデータベースと、

前記検索データベースに登録されていない文書ファイルを検索条件に指定したネットワークからの検索要求によって受信したファイル内容からテキスト文を抽出するテキスト抽出処理部と、

前記テキスト文の形態素解析により名詞を抽出する形態素回析部と、

前記名詞の中から重要語を抽出して論理和でつなげたキーワードを生成するキーワード生成部と、

前記キーワードによるデータベースの検索で類似する文書を検索して要求元に検索結果を通知する検索実行部と、

を備えたことを特徴とする文書情報検索装置。

【請求項７】

請求項６記載の文書情報検索装置に於いて、前記キーワード生成部は検索要求に伴って受信したファイルから抽出したプロパティ情報を前記キーワードに含めて検索することを特徴とする文書情報検索装置。

【請求項８】

ネットワークを経由した検索要求に基づいて文書情報を検索して応答する文書情報検索方法に於いて、

検索対象文書から抽出した重要語を列挙したインデックス情報を文書毎にデータベースに保存し、

検索要求元で検索条件にファイルを指定した場合に、指定したファイル内容を検索要求と共にネットワークを経由して検索先に送信し、

検索側で、検索要求に伴って受信したファイル内容からテキスト文を抽出すると共にテキスト文の形態素解析により名詞を抽出し、次に名詞の中から重要語を抽出して論理和でつなげたキーワードを生成し、該キーワードによるデータベースの検索で類似する文書を検索して検索結果を応答することを特徴とする文書情報検索方法。

【請求項９】

請求項８記載の文書情報検索方法に於いて、検索要求に伴って受信したファイルから抽出したプロパティ情報を前記キーワードに含めて検索することを特徴とする文書情報検索方法。

【請求項１０】

文書ファイルを検索条件に指定した検索要求を受信するステップと、

検索要求に伴って受信したファイル内容からテキスト文を抽出するステップと

テキスト文の形態素解析により名詞を抽出するステップと、

名詞の中から重要語を抽出して論理和でつなげたキーワードを生成するステップと、

前記キーワードによるデータベースの検索で類似する文書を検索して要求元に検索結果を通知するステップと、

を備えた検索プログラムを格納したコンピュータ可読の記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、大量の文書データの中から必要な文書を迅速に探し出すための文書情報検索装置、方法及び文書情報検索プログラムを格納したコンピュータ可読の記録媒体に関し、特に、文書ファイルそのものを検索条件に指定するという簡単な操作で内容が類似する文書を捜し出す文書情報検索装置、方法及び文書情報検索プログラムを格納したコンピュータ可読の記録媒体に関する。

【０００２】

【従来の技術】

従来、ネットワーク環境を利用した文書管理システムにあっては、インタネットやイーサネット上に存在する大量の文書データから必要な文書を検索してすばやく参照することのできる文書情報検索装置を提供している。

【０００３】

この場合の文書検索は、ユーザが必要とする文書に含まれていると思われる１又は複数の適当な単語や文字列をキーワードとして指定し、この指定したキーワードの単語を含む文書を検索データベースから検索し、文書一覧を検索結果として表示する。

【０００４】

この文書情報検索装置にあっては、ネットワーク上に存在する検索対象文書について、その内容から重要語を抽出して列挙したインデックスを文書毎に作成し

て検索データベースに保存している。そしてユーザからキーワードを指定した検索要求があれば、検索データベースのインデックスを検索して文書一覧の検索結果を出すようにしている。

【０００５】

更に、従来の文書情報検索装置は、ユーザがキーワード指定で検索した文書一覧の中から必要と思われる文書を検索した後、文書一覧の中から選んだ文書について類似文書検索を指定すると、検索文書の中に出現する頻度の高い用語が自動的に抽出され、前回実行された検索条件に論理和の条件で付加され、類似文書の検索を行うことができる。

【０００６】

【発明が解決しようとする課題】

しかしながら、ユーザが電子メールやインターネットで、興味ある文書を入手し、この文書に類似した内容の文書を検索したい場合、現状では入手した文書に含まれている単語や文字列を選んでキーワードにいちいち指定し、まず検索結果として文書一覧を得る。次に、検索した文書一覧の中から文書を選択して類似文書検索を指定して類似文書の検索を行わなければならない。

【０００７】

即ち、電子メールやインターネットで入手した文書の類似検索を行おうとしても、従来の文書情報検索装置は、既に検索データベースに登録されている文書しか、文書を検索条件に指定した類似文書の検索はできず、ユーザが電子メールやインターネットで入手した文書を検索条件に使って直接的に類似文書の検索を行うことができない。

【０００８】

このためユーザが電子メールやインターネットで入手した文書の中から、文書検索に必要と思われるキーワードを選んで検索条件として入力する必要があり、キーワードが多くある場合は入力に手間がかかる。またキーワードの指定が十分でないと検索漏れを生じ、期待した検索結果が得られない場合がある。

【０００９】

更に文書一覧として得られる検索数が膨大となることもあり、文書一覧から関連すると思われる文書を開いて必要な文書を捜し出す大変な手間がかかる場合がある。

【００１０】

本発明は、検索データベースに登録されていない文書に類似した文書の検索を簡単な操作ですばやくできる文書情報検索装置、方法及び文書情報検索プログラムを格納したコンピュータ可読の記録媒体を提供することを目的とする。

【００１１】

【課題を解決するための手段】

図１は本発明の原理説明図である。本発明は、クライアント１２等からのネットワークを経由した検索要求に基づいてサーバ１０等の検索側で文書情報を検索して応答する文書情報検索装置であって、クライアント１２等の要求元に、検索条件にファイルを指定した場合に、指定したファイル内容をネットワークを経由して送信する検索条件指定部２６を設け、サーバ１０等の検索側に、検索条件指定部２６から送信されたファイル内容からキーワードを生成して類似文書を検索する検索マシン２０を設けたことを特徴とする。

【００１２】

このため電子メールやインターネット等で興味のある内容を含む文書を入手し、この文書に類似した内容の文書を検索したい場合等に、文書の指定によりアップロードされたファイルを検索条件に指定することで、内容が類似する文書を検索することができる。このためデータベース登録されていない文書であっても自由に検索条件として指定することができ、手間のかかる文書内容に基づいたキーワードの入力を不要とし、簡単且つ迅速に類似文書を探し出すことができる。

【００１３】

検索要求元の検索条件指定部２６は、指定されたファイル内容の先頭ファイル部分を送信する。通常、文書検索に必要な重要なキーワードは文書の先頭部分に多く存在することから、ファイル内容の先頭部分だけ、例えば先頭の１ＫＢ部分を検索条件として送信する。また検索条件に使用する文書ファイルのサイズは様

々であることから、検索条件として送信するファイル容量を決めることで、通信負荷と検索側の処理を軽減する。

【００１４】

検索条件指定部２６は検索条件として指定するファイルにＨＴＭＬファイル及びエクセルファイルを含む。勿論、これ以外のファイル形式であっても、テキスト文書の抽出が可能なファイルであれば、任意のファイル形式のものを含む。

【００１５】

サーバ１０側の検索マシン２０には、検索対象文書から抽出した重要語を列挙したインデックス情報を文書毎に保存したデータベース２２が設けられる。また検索マシン２０のファイル指定検索部３０は、検索要求に伴って受信したファイル内容からテキスト文を抽出するテキスト抽出処理部３６、テキスト文の形態素解析により名詞を抽出する形態素回析部３８、名詞の中から重要語を抽出して論理和でつなげたキーワードを生成するキーワード生成部４０、及びキーワードによる検索データベース２２の検索で類似する文書を検索してクライアントに検索結果を通知する検索実行部４２を備える。

【００１６】

キーワード生成部４０は、各名詞が検索データベース２２に格納した検索文書毎のインデックス中の何文書に出現するかの出現数Ｈをカウントし、所定範囲の出現数Ｈをもつ上位の所定数の単語を選択してキーワードを生成する。

【００１７】

キーワード生成部４０は、インデックス中の文書数Ｎとした場合、例えば出現数Ｈが

$$2N/3 \geqq H \geqq 1$$

の範囲の出現数をもつ上位の１０個の単語を選択してキーワードを生成する。これによりデータベースのインデックスに登録している既存文書の類似検索に必要な重要語を絞り込み、類似検索の精度を高める。

【００１８】

更にキーワード生成部４０は、検索要求に伴って受信したファイルから抽出したプロパティ情報をキーワードに含めて検索させる。この場合のプロパティ情報

は、検索要求に伴って受信したファイルの作成者、文書タイトル等である。このように検索条件に、ファイルのプロパティ情報を加えることで、例えば作成者等を特定したい場合の類似文書の絞り込みが適切にできる。

【００１９】

検索要求元の検索条件指定部２６はクライアント１２のWWWブラウザ１６で提供され、WWWブラウザ１６の検索要求画面で指定したファイル内容をネットワークを介してWWWサーバ１８に送信して検索マシン２０に引き渡す。

【００２０】

本発明は、またサーバ等の検索側の文書情報検索装置となる検索マシン２０を提供する。この検索マシン２０としての文書情報検索装置は、検索対象文書から抽出した重要語を列挙したインデックス情報を文書毎に保存している検索データベース２２、文書ファイルを検索条件に指定したネットワークからの検索要求によって受信したファイル内容からテキスト文を抽出するテキスト抽出処理部３６、テキスト文の形態素解析により名詞を抽出する形態素回析部２８、名詞の中から重要語を抽出して論理和でつなげたキーワードを生成するキーワード生成部４０、及びキーワードによる検索データベースの検索で類似する文書を検索して要求元に検索結果を通知する検索実行部４２を備える。

【００２１】

本発明は、クライアント等の検索要求元からのネットワークを経由した検索要求に基づいてサーバ等の検索マシン側で文書情報を検索して応答する文書情報検索方法を提供する。この文書情報検索方法は、

検索対象文書から抽出した重要語を列挙したインデックス情報を文書毎にサーバの検索データベースに保存し；

文書ファイルを検索条件に指定した場合に、指定したファイル内容を検索要求と共にネットワークを経由して検索側に送信し；

検索側で、検索要求に伴って受信したファイル内容からテキスト文を抽出すると共にテキスト文の形態素解析により名詞を抽出し、次に名詞の中から重要語を抽出して論理和でつなげたキーワードを生成し、該キーワードによる検索データベースの検索で類似する文書を検索してクライアントに検索結果を通知すること

を特徴とする。この文書情報検索方法の詳細は装置構成と基本的に同じになる。

【００２２】

更に、本発明は、文書情報検索プログラムを格納したコンピュータ可読の記録媒体を提供するもので、この文書情報検索プログラムは、文書ファイルを検索条件に指定した検索要求を受信するステップと、検索要求に伴って受信したファイル内容からテキスト文を抽出するステップと、テキスト文の形態素解析により名詞を抽出するステップと、名詞の中から重要語を抽出して論理和でつなげたキーワードを生成するステップと、キーワードによるデータベースの検索で類似する文書を検索して要求元に検索結果を通知するステップとを備える。

【００２３】

【発明の実施の形態】

図２は、本発明による文書情報検索装置のシステム構成であり、インターネットやイーサネットを利用したサーバクライアント型の検索システムとして構築した場合を例にとっている。

【００２４】

図２において、サーバ１０に対しては、ユーザ側のクライアント１２がインターネット／イントラネット１４を介して接続される。クライアント１２には検索用のＷＷＷブラウザ１６が設けられており、このＷＷＷブラウザ１６を利用してサーバ１０に対し文書情報の検索要求を行い、サーバ１０側の検索結果を表示する。

【００２５】

サーバ１０には、ＷＷＷサーバ１８、検索マシン２０、文書データベース２４が設けられている。検索マシン２０には検索データベース２２が格納されている。また文書データベース２４には検索対象文書２５が格納されている。更にＷＷＷサーバ１８に対して外部の文書管理サーバ４４，４８が接続され、この文書管理サーバ４４，４８にも文書データベース４６，５０が設けられており、それぞれ検索対象文書２５を格納している。

【００２６】

サーバ１０に設けているWWWサーバ１８は、ブラウザ１６からの検索要求を受信して検索マシン２０に対し検索を依頼する。また検索マシン２０から返ってきた検索結果をブラウザ１６に返して表示させる。

【００２７】

検索データベース２２は、全文検索を高速に処理するために、検索対象となる文書に記述されている重要な単語の集合で作られたインデックスを管理する保管庫として機能する。このインデックスには文書の文書名やその保管場所が記録されており、ブラウザ１６から検索要求を受けた際には、検索データベース２２のインデックスを対象に検索マシン２０が検索処理を実行する。

【００２８】

文書データベース２４には、文書管理サーバ４４，４８から収集した検索対象文書２５が格納されており、この文書データベース検索対象文書２５を対象に検索データベース２２のインデックスが作成されている。

【００２９】

このようなサーバクライアント型の検索システムにあっては、クライアント１２のブラウザ１６を使用して、ユーザが指定した検索条件をインターネット／イントラネット１４を経由してサーバ１０側のWWWサーバ１８に送る。WWWサーバ１８で受信された検索要求に含まれる指定された検索条件が、WWWサーバ１８から検索マシン２０に送られる。

【００３０】

検索マシン２０は検索データベース２２から検索条件にあった文書を検索し、検索結果をWWWサーバ１８に通知する。WWWサーバ１８は検索マシン２０からの検索結果をクライアント１２のブラウザ１６に送って表示させる。

【００３１】

ユーザはブラウザ１６で処理された検索結果を見て、検索結果に記述されたリンクを選択することで、選択された文書の中からユーザが希望する検索対象文書２５をWWWサーバ１８経由でアップロードして内容を見ることができる。

【００３２】

図３は図２の検索システムにおける機能構成のブロック図である。まずユーザ

側となるWWWブラウザ１６には検索条件指定部２６が設けられている。本発明の検索条件指定部２６は、検索条件としてユーザがインターネットや電子メールなどで入手した文書ファイルを直接、検索条件として指定し、指定したファイル内容をインターネット／イントラネット１４経由でWWWサーバ１８を経由して検索マシン２０のファイル指定検索部３０に送信する。

【００３３】

また検索条件指定部２６は、本発明で新たに提供されるファイル指定の検索条件とする以外に、

（１）キーワード検索、

（２）文書のタイトル、作成者、本文ごとにキーワードを指定して検索する詳細検索、

（３）日常的な言葉や文章を入力することにより本文内容を関連する文書を検索する文章検索、更に、

（４）検索データベース２２に登録済みの既存文書を検索条件に使用した類似文書検索、

などの検索条件の指定も可能である。

【００３４】

WWWサーバ１８側に設けられた検索マシン２０には、検索データベース作成部２８、文書検索部３０及び文書参照部３２が設けられている。検索データベース作成部２８は検索データベース２２にインデックスを作成して登録する。

【００３５】

即ち検索データベース作成部２８は、文書データベース２４に収集されて保存されている検索対象文書２５の１つ１つについて、検索対象文書２５に記述されている重要語を抽出し、抽出された単語の集合で構成されたインデックスを作成して保存する。もちろん、このインデックスには検索対象文書の文書名や保管場所などが併せて記録されている。

【００３６】

文書検索部３０は、WWWブラウザ１６の検索条件指定部２６から送信された検索条件としてファイルを指定した際のファイル内容からキーワードを生成し、

検索データベース２２のインデックスに含まれている重要単語の集合との検索照合を行い、ＷＷＷブラウザ１６で検索条件として指定したファイルの文書に類似する文書を検索し、検索結果をＷＷＷサーバ１８からＷＷＷブラウザ１６に返して表示させる。

【００３７】

文書参照部３２は、ＷＷＷブラウザ１６で送出された検索結果としての文書一覧から参照したい文書を選択すると、ＷＷＷサーバ１８を介して文書参照部３２に通知されると、文書データベース２４の中から要求された参照文書を取り出してＷＷＷブラウザ１６に返す。

【００３８】

図４は、図３の検索マシン２０に設けた本発明の文書検索部３０の機能構成の詳細である。

【００３９】

図４において、文書検索部３０には、検索指定ファイル格納部３４、テキスト抽出処理部３６、形態素解析部３８、キーワード作成部４０及び検索実行部４２が設けられている。また検索データベース２２内には、図３の検索データベース作成部２８で作成された文書データベース２４内の検索対象文書２５のそれぞれの重要単語の集合、文書名、保管場所などで構成されたインデックス５２が格納されている。

【００４０】

文書検索部３０の検索指定ファイル格納部３４には、図３のＷＷＷブラウザ１６における検索条件指定部２６のファイル指定により送信されたファイル内容が格納される。

【００４１】

ここでＷＷＷブラウザ１６側からのファイル内容の転送は、検索条件として指定した文書ファイルの先頭ファイル部分、例えば先頭の１ＫＢを切り出してＷＷＷサーバ１８側に検索要求と共に送信する。

【００４２】

このように検索条件として送信するファイル容量を例えば１ＫＢというように

固定容量とすることで、検索条件として指定している文書ファイルのサイズの大小に関わらず、検索マシン２０側に対する文書内容の転送負荷を一定にし、また検索マシン２０におけるファイル指定部検索部３０による検索処理の安定化と迅速化を図る。

【００４３】

テキスト抽出処理部３６は、検索指定ファイル格納部３４に格納された検索条件として指定されたファイル内容からテキスト文書を抽出する。ＷＷＷブラウザ１６における検索条件として指定される文書ファイルの形式としては、電子メールのテキストファイル、インターネットにおけるＨＴＭＬファイル、更には集計リストのエクセルファイルなどの様々なファイル形式があることから、これらのファイル形式の相違に対して検索機能を提供可能とするため、各種の形式の文書ファイルの中からテキスト抽出処理部３６によりテキスト文書のみを抽出して検索条件に使用するようにしている。

【００４４】

続いて設けた形態素解析部３８は、抽出されたテキスト文書の中に含まれる名詞を形態素解析を用いて抽出する。形態素解析部３８で抽出された文書内容の中の名詞はキーワード作成部４０に送られ、キーワード作成部４０においては重要な名詞をキーワード作成のために抽出する。

【００４５】

キーワード作成部４０における重要語の抽出は、まず各名詞が検索データベース２２のインデックス５２の中に登録している文書数Ｎの内の何文書で出現するかの出現数Ｈのカウントを行う。

【００４６】

そして、インデックス５２中における文書出現数Ｈが求められたならば、出現数Ｈが予め定めた範囲内、例えば

$$(2N/3) \geqq H \geqq 1$$

となる出現数の単語を選択する。このように選択された単語の内の出現数Ｈが大きい上位１０個の単語をキーワード作成のために選択する。そして選択した重要単語１０個を論理和で繋げたクエリ式を作成して検索実行部４２に提供する。

【００４７】

検索実行部４２はキーワード作成部４０から与えられたクエリ式に基づいて検索データベース２２のインデックス５２との検索照合を行い、所定の類似度を満たすインデックスを検索結果として抽出し、検索結果をＷＷＷサーバ１８によりＷＷＷブラウザ１６側に送信し、検索結果の文書一覧の形でユーザに参照できるようにする。

【００４８】

更に文書検索部３０にあっては、検索指定ファイル格納部３４に格納された検索条件として指定されたファイルのプロパティ情報を利用した文書検索もできる。このためＷＷＷブラウザ１６の検索条件指定部２６は、検索条件として文書ファイルを指定した際に、指定した文書ファイルのプロパティ情報を抽出し、検索条件として指定した文書の先頭ファイル部分、例えば先頭ファイル部分１ＫＢと共にプロパティ情報を検索マシン２０側に送信する。

【００４９】

図１４の文書検索部３０にあっては、ファイル内容からのテキスト文の抽出、形態素解析による名詞抽出、名詞について重要語の選択によるキーワード作成に加え、検索指定ファイル格納部３４に格納されているファイル内容に付加されたプロパティ情報から例えば作成日や作成者、題名などを抽出し、キーワード作成部４０でプロパティ情報をキーワードに含め、検索実行部４２で検索データベース２２のインデックス５２の検索を行う。

【００５０】

図５は、図３の検索マシン２０に設けている検索データベース作成部２８によるインデックス作成処理の説明図である。この検索データベース作成部２８にあっては、ロボット５４が外部の文書データベース４６，５０から文書６６を収集してテンポラリファイル６２に格納し、同時に収集文書リストファイル６４に収集した文書６６のリストを加える。

【００５１】

続いてロボット５４はテキスト抽出部５６に処理を渡し、テキスト抽出部５６は収集文書リストファイル６４から収集文書６６を取り出し、抽出テキストファ

イル６８に格納する。

【００５２】

次に重要語抽出部５８に処理を渡し、重要語抽出部５８は抽出テキストファイル６８の該当テキスト文書の中から形態素解析により名詞を抽出し、名詞についてそれぞれ出現頻度をカウントし、例えば出現頻度の高い単語の上位１０個を重要語として抽出して重要語ファイル７０に格納する。

【００５３】

次にインデックス作成部６０に処理を渡し、インデックス作成部６０は重要語ファイル７０から、その文書について例えば上位１０個の重要語の集合を取り出し、更に文書名と保管場所を加えたインデックスを作成し、検索データベース２２にインデックス情報として保存する。

【００５４】

図６は、図３のＷＷＷブラウザ１６による検索条件の指定と検索結果の表示を行うブラウザ処理のフローチャートである。ユーザがＷＷＷブラウザ１６の検索機能を開くと、ステップＳ１で検索画面が表示され、この検索画面を表示して、ステップＳ２で文書ファイルを指定した検索条件の指定操作を行う。

【００５５】

続いてステップＳ３で検索起動の有無をチェックしており、検索起動を判別すると、ステップＳ４でファイル指定検索か否かチェックする。ファイル指定検索であればステップＳ５に進み、ユーザが指定したファイルを読み出し、ステップＳ６で指定ファイルの先頭１ＫＢを検索要求メッセージと共にサーバに送信する。

【００５６】

ファイル指定検索でなければ、ステップＳ７で、それ以外の検索例えばキーワード検索に対応した検索要求メッセージをサーバに送信する。ステップＳ６で指定ファイルの先頭部分をサーバに送信すると、ステップＳ８で検索結果の受信待ちとなる。

【００５７】

ステップＳ８でサーバから検索結果が受信されると、ステップＳ９に進み、検

索結果の表示操作処理を行ってユーザは検索内容を見る。このようなステップＳ１～Ｓ９の処理を、ステップＳ１０で検索画面を閉じる検索終了指示があるまで繰り返す。

【００５８】

図７は、図６のブラウザ処理において検索条件として文書ファイルを指定した場合の具体的な手順と画面の様子を表わしている。

【００５９】

図７において、まずユーザは検索条件に指定しようとする文書ファイル７２を例えばインターネットから取得している。そしてユーザは文書ファイル７２の内容を見て、この文書ファイル７２に類似する文書検索を行うため、文書ファイル７２の内容を予め指定したファイル、例えばファイル「ｎｅｗｓ．ｔｘｔ」に保存する。

【００６０】

続いてユーザはキーワード入力画面７４を開く。キーワード入力画面７４にはキーワード入力部７６、ファイル指定部７８、参照ボタン８０及び検索実行ボタン８２が設けられている。そこで、ユーザがキーワード入力画面７４の参照ボタン８０を押すことでファイル選択ダイアログ８４を表示する。

【００６１】

このファイル選択ダイアログ８４の中には、検索条件として指定したい文書ファイル７２が保存されていることから、ファイル名「ｎｅｗｓ．ｔｘｔ」をマウスクリックして選択すると、キーワード入力画面７４のファイル指定部７８に選択したファイル名「ｎｅｗｓ．ｔｘｔ」が設定される。

【００６２】

このようにしてファイル指定部７８によるファイル指定が済んだならば、検索実行ボタン８２を押すことで、検索条件として指定された文書ファイル「ｎｅｗｓ．ｔｘｔ」の文書内容の先頭１ＫＢが検索要求と共にサーバに対し送信される。

【００６３】

図８は、図４の文書検索部３０によって実現されるサーバ検索処理のフローチ

ャートである。このサーバ検索処理は、ステップＳ１で検索条件として指定された文書ファイルを読み込み、ステップＳ２で文書ファイルからテキスト文書の抽出処理を行う。次にステップＳ３で、抽出したテキスト文書の内容について形態素解析を用いて名詞を抽出する。次にステップＳ４で、名詞として抽出した各単語が検索データベース２２に設けているインデックス５２の中の文書数Ｎの内の何文書に出現するかの出現数Ｈのカウント処理を行う。

【００６４】

各単語のインデックス中の出現数Ｈがカウントできたならば、ステップＳ５で出現数Ｈが（２Ｎ／３）以下で１以上となる範囲の単語をまず選択し、この選択した単語のうち出現数Ｈが大きい上位１０個の単語をキーワードに使用する重要語として選択する。続いてステップＳ６で、重要語として選択した１０個の単語を論理和で繋げたクエリ式を生成する。

【００６５】

そしてステップＳ７で、検索キーワードとして生成されたクエリ式による検索データベースのインデックスの検索を行い、生成したキーワードに対し所定の類似度を持つインデックスの内容を検索文書として一覧表にまとめ、ステップＳ８で検索結果をブラウザに送信する。

【００６６】

図９は、図８のステップＳ２のテキスト抽出処理の詳細である。このテキスト抽出処理にあっては、ステップＳ１で文書ファイルの拡張子を解読する。ファイル拡張子からステップＳ２でＨＴＭＬ文書であることが認識されると、ステップＳ３に進み、ＨＴＭＬ文書におけるボディタグ内のデータをテキストデータ本文として抽出し、タグデータは取り除く。

【００６７】

例えば図１０（Ａ）のようなＨＴＭＬファイルを例にとると、＜　＞で挟まれたボディ単語の中のデータをテキストデータ本文として取り出して、このタグデータは取り除くことで、図１０（Ｂ）のような抽出テキスト文書が得られる。

【００６８】

次にステップＳ４で、ＯＳで管理しているファイルのプロパティ情報を獲得す

る。このプロパティ情報は、例えばファイル所有者や文書タイプなどを含んでいる。

【００６９】

図１１は、インターネットから入手した文書ファイルのプロパティ情報の例であり、このプロパティ情報にあっては文書タイトル「文書管理システムについて」や作成日、変更日などが存在し、これらのプロパティデータをキーワード生成のために獲得する。

【００７０】

一方、ステップＳ２でＨＴＭＬ文書ではなく例えばエクセル文書などであった場合には、ステップＳ５で文書ライブラリにファイルを渡し、テキストデータを獲得する。続いてステップＳ６で、プロパティ情報獲得関数により文書ごとに設定されているファイルプロパティ情報例えば作成者や文書タイトルなどを獲得する。

【００７１】

図１２は本発明で検索条件として指定するＨＴＭＬファイル以外のファイルとしてエクセルファイルを示している。この図１２のエクセルファイルについて、文書ライブラリに渡してテキストデータを獲得すると、図１３の抽出テキスト文書に示すようなエクセル文書中に書き込まれているテキスト文書を抽出した結果が得られる。

【００７２】

このようなテキスト抽出処理で得られたＨＴＭＬ文書やエクセル文書からのテキスト文書、更にはプロパティ情報から得られたテキスト文書をひとまとめにして、図８のステップＳ３で形態素解析を用いて名詞を抽出し、ステップＳ４，Ｓ５で、データベースのインデックスの参照で重要語の上位１０個をキーワードに選択してクエリ式を作り、データベースのインデックス検索を行って検索結果を得ることができる。

【００７３】

尚、図９のテキスト抽出処理におけるステップＳ４，Ｓ６のプロパティ情報の獲得は、ＷＷＷブラウザ１６におけるユーザ側の指定によってプロパティ情報を

使用するか否かの選択が可能であり、プロパティ情報を使うか否かは検索結果をどの程度絞り込むかのユーザ判断に依存する。

【００７４】

本発明はまた、図４の検索マシン２０に文書検索部３０の処理機能を実行する文書情報検索プログラムを記録したコンピュータ読取り可能な記録媒体を提供する。この記録媒体の実施形態としては、ＣＤ－ＲＯＭやフロッピディスクなどのリムーバブルな可搬型記録媒体、回線によりプログラムを提供するプログラム提供者の記憶装置、更にプログラムをインストールした処理装置のＲＡＭやハードディスクなどのメモリ装置を含む。

【００７５】

また記録媒体によって提供された図４の文書検索部３０の機能を実現する文書情報検索プログラム、具体的には図８及び図９のフローチャートの処理を実行するステップを備えた文書情報検索プログラムは、サーバなどの処理装置にローディングされ、その主メモリ上で実行される。

【００７６】

またサーバ側にローディングされた本発明の文書情報検索プログラムは、クライアント側からサービス要求を受けると、クライアント１２側にファイル指定による検索条件の指定を行うＷＷＷブラウザ機能をアップロードし、ユーザによる検索システムの利用を可能とする。

【００７７】

尚、上記の実施形態はサーバクライアント型の検索システムを例にとるものであったが、本発明はこれに限定されず、ホスト端末型や適宜のシステム形態をとることができる。また本発明は上記の実施形態に限定されず、その目的と利点を損なわない適宜の変形を含む。更にまた本発明は上記の実施形態に示した数値による限定は受けない。

（付記）

（付記１）

ネットワークを経由した検索要求に基づいて文書情報を検索して応答する文書情

報検索装置に於いて、
検索要求元に、検索条件としてファイルを指定し、指定したファイル内容をネットワークを経由して送信する検索条件指定部を設け、
検索側に、前記検索条件指定部から送信されたファイル内容からキーワードを生成してデータベースから類似文書を検索する文書検索部を設けたことを特徴とする文書情報検索装置。（１）

（付記２）
付記１記載の文書情報検索装置に於いて、前記検索条件指定部は、指定されたファイル内容の先頭ファイル部分を送信することを特徴する文書情報検索装置。

【００７８】

（付記３）
付記１記載の文書情報検索装置に於いて、前記検索条件指定部は検索条件として指定するファイルにＨＴＭＬファイル及びエクセルファイルを含むことを特徴とする文書情報検索装置。

【００７９】

（付記４）
付記１記載の文書情報検索装置に於いて、
前記データベースは、検索対象文書から抽出した重要語を列挙したインデックス情報を文書毎に保存し、
サーバの文書検索部は、
検索要求に伴って受信したファイル内容からテキスト文を抽出するテキスト抽出処理部と、
前記テキスト文の形態素解析により名詞を抽出する形態素回析部と、
前記名詞の中から重要語を抽出して論理和でつなげたキーワードを生成するキーワード生成部と、
前記キーワードによる検索データベースの検索で類似する文書を検索してクライアントに検索結果を通知する検索実行部と、
を備えたことを特徴とする文書情報検索装置。（２）

（付記５）

付記４記載の文書情報検索装置に於いて、前記キーワード生成部は、各名詞が前記文書データベースに格納した検索文書毎のインデックス中の何文書に出現するかの出現数をカウントし、所定範囲の出現数をもつ上位の所定数の単語を選択してキーワードを生成することを特徴とする文書情報検索装置。（３）

（付記６）

付記５記載の文書情報検索装置に於いて、前記キーワード生成部は、インデックス中の文書数Ｎとした場合、出現数Ｈが２Ｎ／３≧Ｈ≧１の範囲の出現数をもつ上位の１０個の単語を選択してキーワードを生成することを特徴とする文書情報検索装置。（４）

（付記７）

付記３記載の文書情報検索装置に於いて、前記キーワード生成部は検索要求に伴って受信したファイルから抽出したプロパティ情報を前記キーワードに含めて検索させることを特徴とする文書情報検索装置。（５）

（付記８）

付記７記載の文書情報検索装置に於いて、前記プロパティ情報は、検索要求に伴って受信したファイルの作成者、文書タイトル等であることを特徴とする文書情報検索装置。

【００８０】

（付記９）

付記１記載の文書情報検索装置に於いて、前記検索要求元の検索条件指定部はクライアントのＷＷＷブラウザで提供され、前記ＷＷＷブラウザの検索要求画面で指定したファイル内容をネットワークを介してＷＷＷサーバの検索マシンに送信して前記ファイル指定検索部に引き渡すことを特徴とする文書情報検索装置。

【００８１】

（付記１０）

検索対象文書から抽出した重要語を列挙したインデックス情報を文書毎に保存しているデータベースと、

文書ファイルを検索条件に指定したネットワークからの検索要求によって受信したファイル内容からテキスト文を抽出するテキスト抽出処理部と、

前記テキスト文の形態素解析により名詞を抽出する形態素回析部と、
前記名詞の中から重要語を抽出して論理和でつなげたキーワードを生成するキーワード生成部と、
前記キーワードによるデータベースの検索で類似する文書を検索して要求元に検索結果を通知する検索実行部と、
を備えたことを特徴とする文書情報検索装置。（６）

（付記１１）
付記１０記載の文書情報検索装置に於いて、前記キーワード生成部は、各名詞が前記文書データベースに格納した検索文書毎のインデックス中の何文書に出現するかの出現数をカウントし、所定範囲の出現数をもつ上位の所定数の単語を選択してキーワードを生成することを特徴とする文書情報検索装置。

【００８２】

（付記１２）
付記１０記載の文書情報検索装置に於いて、前記データベースにインデックス情報と共に検索対象文書から抽出したプロパティ情報を保存し、前記キーワード生成部は検索要求に伴って受信したファイルから抽出したプロパティ情報を前記キーワードに含めて検索することを特徴とする文書情報検索装置。（７）

（付記１３）
ネットワークを経由した検索要求に基づいて文書情報を検索して応答する文書情報検索方法に於いて、
検索対象文書から抽出した重要語を列挙したインデックス情報を文書毎にデータベースに保存し、
前記検索要求元で検索条件にファイルを指定した場合に、指定したファイル内容を検索要求と共にネットワークを経由してサーバに送信し、
検索側で、検索要求に伴って受信したファイル内容からテキスト文を抽出すると共にテキスト文の形態素解析により名詞を抽出し、次に名詞の中から重要語を抽出して論理和でつなげたキーワードを生成し、該キーワードによるデータベースの検索で類似する文書を検索して検索結果を応答することを特徴とする文書情報検索方法。（８）

（付記１４）

付記１３記載の文書情報検索方法に於いて、前記キーワードの生成として、各名詞が前記データベースに格納した文書毎のインデックス中の何文書に出現するかの出現数をカウントし、所定範囲の出現数をもつ上位の所定数の単語を選択してキーワードを生成することを特徴とする文書情報検索方法。

【００８３】

（付記１５）

付記１４記載の文書情報検索方法に於いて、検索要求に伴って受信したファイルから抽出したプロパティ情報を前記キーワードに含めて検索することを特徴とする文書情報検索方法。（９）

（付記１６）

文書ファイルを検索条件に指定した検索要求を受信するステップと、

検索要求に伴って受信したファイル内容からテキスト文を抽出するステップと、

テキスト文の形態素解析により名詞を抽出するステップと、

名詞の中から重要語を抽出して論理和でつなげたキーワードを生成するステップと、

前記キーワードによるデータベースの検索で類似する文書を検索して要求元に検索結果を通知するステップと、

を備えた検索プログラムを格納したコンピュータ可読の記録媒体。（１０）

（付記１７）

付記１６記載の記録媒体に於いて、前記文書情報検索プログラムのキーワードを生成するステップは、各名詞が前記データベースに格納した文書毎のインデックス中の何文書に出現するかの出現数をカウントし、所定範囲の出現数をもつ上位の所定数の単語を選択してキーワードを生成することを特徴とする記録媒体。

【００８４】

（付記１８）

付記１４記載の記録媒体に於いて、前記文書情報検索プログラムは、更に検索要求に伴って受信したファイルから抽出したプロパティ情報を前記キーワードに含めて検索するステップを備えたことを特徴とする記録媒体。

【００８５】

【発明の効果】

以上説明してきたように本発明によれば、ユーザが電子メールやインターネットなどで興味のある内容を含む文書を入手した際に、この文書に類似した内容の文書検索を文書ファイルを直接検索条件として指定することで、内容が類似する文書を簡単且つ素早く検索することができ、手間の掛かる文書内容に基づいたキーワードの入力を不要とし、ユーザによる類似文書の探し出しが極めて効率的に実現できる。

【００８６】

またファイル指定による文書検索に必要なキーワードの生成において、文書内容から重要な単語を抽出する以外に、文書ファイルの持っているプロパティ情報からも重要な単語を抽出してキーワードに含めることで、データベースに登録している既存文書の類似検索の絞り込みが、より適切に行われ、検索の精度を高めることができる。

【図面の簡単な説明】

【図１】

本発明の原理説明図

【図２】

本発明のシステム構成の説明図

【図３】

本発明の機能構成のブロック図

【図４】

本発明による文書検索部のブロック図

【図５】

図３の検索データベース作成部の処理説明図

【図６】

図３のブラウザ処理のフローチャート

【図７】

本発明の検索条件に文書ファイルを指定する検索要求操作の説明図

【図８】

本発明のサーバ検索処理のフローチャート

【図９】

図８のテキスト抽出処理のフローチャート

【図１０】

図８の処理によりＨＴＭＬファイルからのテキスト文書を抽出する説明図

【図１１】

本発明の検索に使用するＨＴＭＬファイルに設けたプロパティ情報の説明図

【図１２】

図８の処理によりテキスト抽出対象とするＥｘｃｅｌ文書の説明図

【図１３】

図１２のＥｘｃｅｌ文書から抽出したテキスト文書の説明図

【符号の説明】

１０：サーバ

１２：クライアント

１４：インターネット／イントラネット

１６：ＷＷＷブラウザ

１８：ＷＷＷサーバ

２０：検索マシン

２２：検索データベース

２４，４６，５０：文書データベース

２５：検索対象文書

２６：検索条件指定部

２８：検索データベース作成部

３０：文書検索部

３２：文書参照部

３４：検索指定ファイル格納部

３６：テキスト抽出処理部

３８：形態素解析部

４０：キーワード作成部

４２：検索実行部

４４，４８：文書管理サーバ

５４：ロボット

５６：テキスト抽出部

５８：重要語抽出部

６０：インデックス作成部

６２：テンポラリファイル

６４：収集文書リストフアファイル

６６：文書

６８：抽出テキストファイル

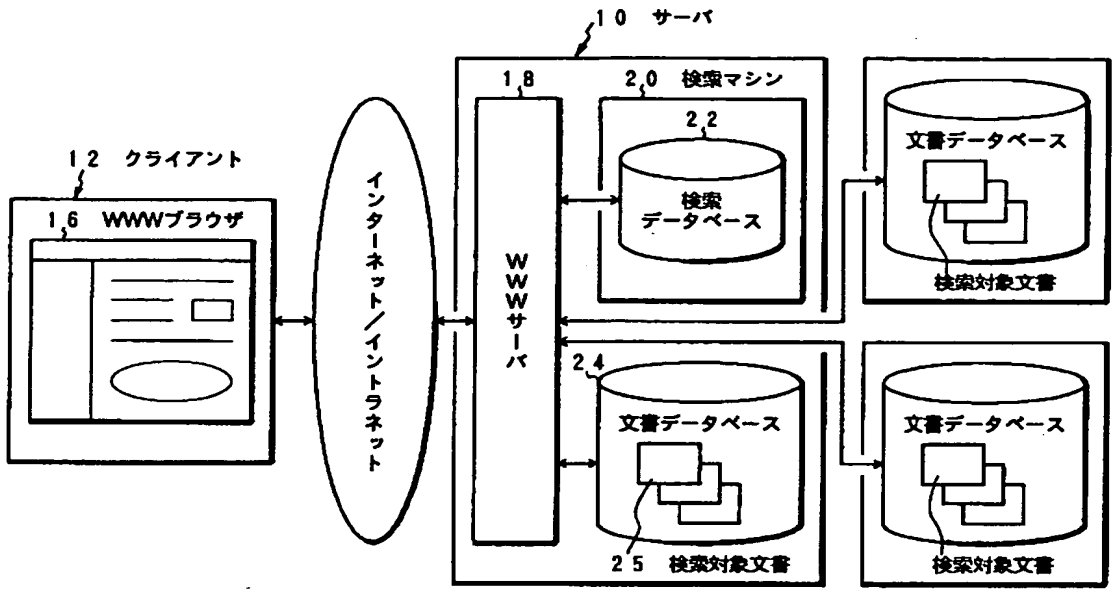
７０：重要語ファイル

【書類名】　　図面

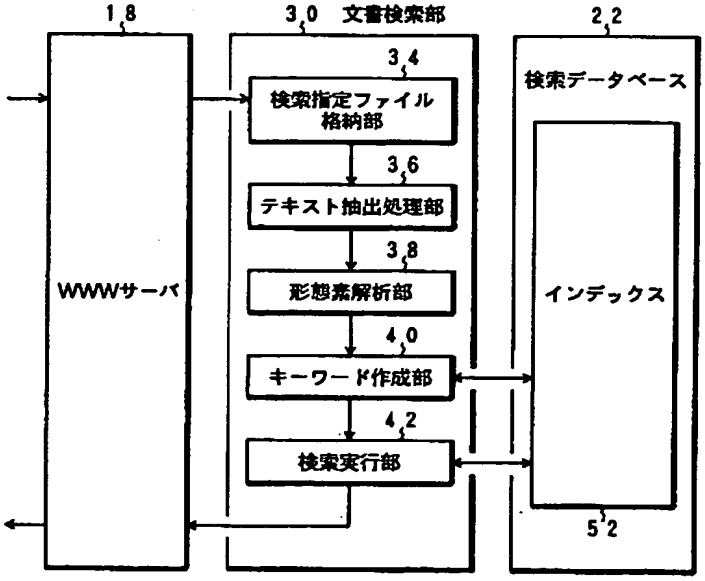
【図１】

本発明の原理説明図

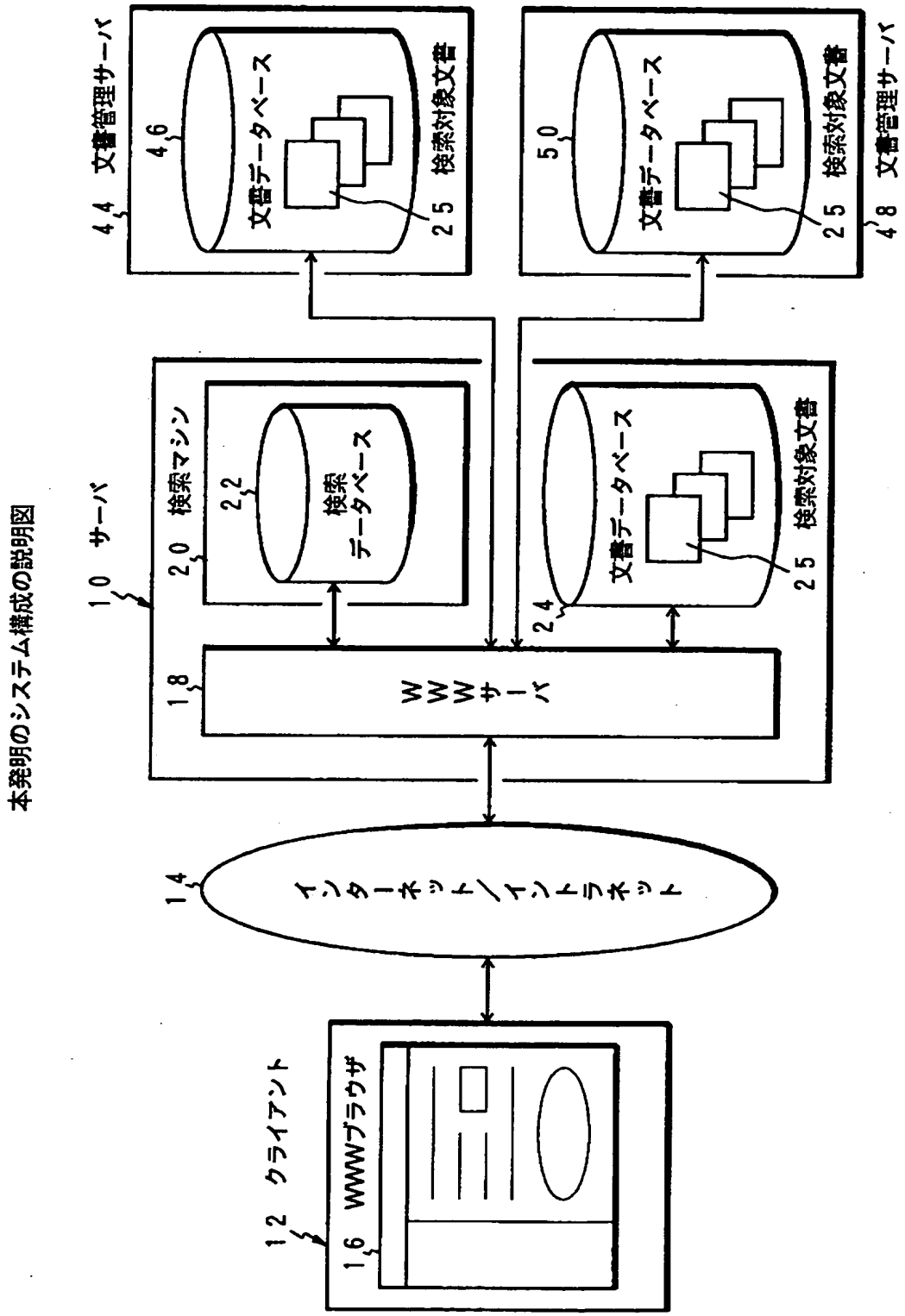
(A)



(B)

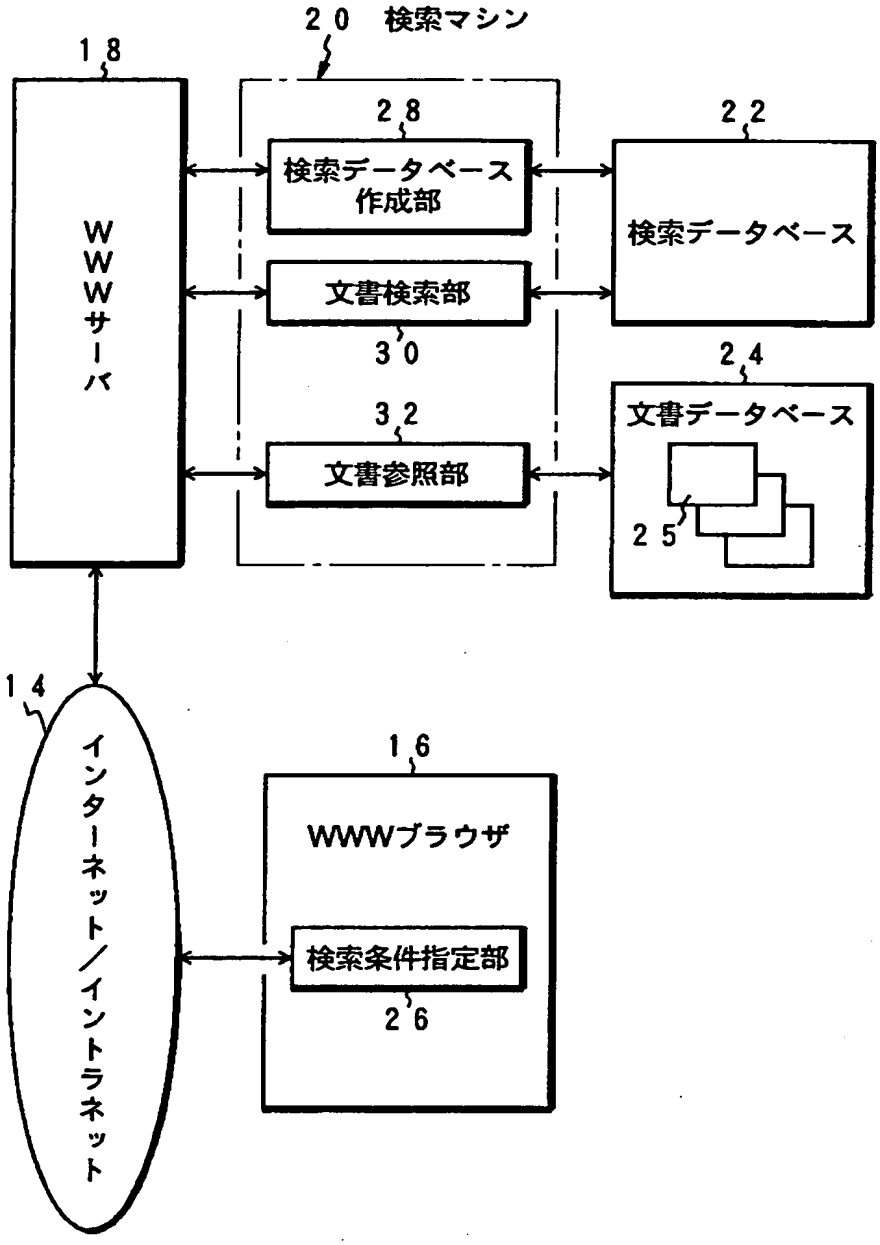


【図２】


本発明のシステム構成の説明図
１０ サーバ
２０ 検索マシン
２２ 検索データベース
１８ WWWサーバ
２４ 文書データベース
２５ 検索対象文書
４４ 文書管理サーバ
４６ 文書データベース
２５ 検索対象文書
５０ 文書データベース
２５ 検索対象文書
４８ 文書管理サーバ
１４ インターネット／イントラネット
１２ クライアント
１６ WWWブラウザ

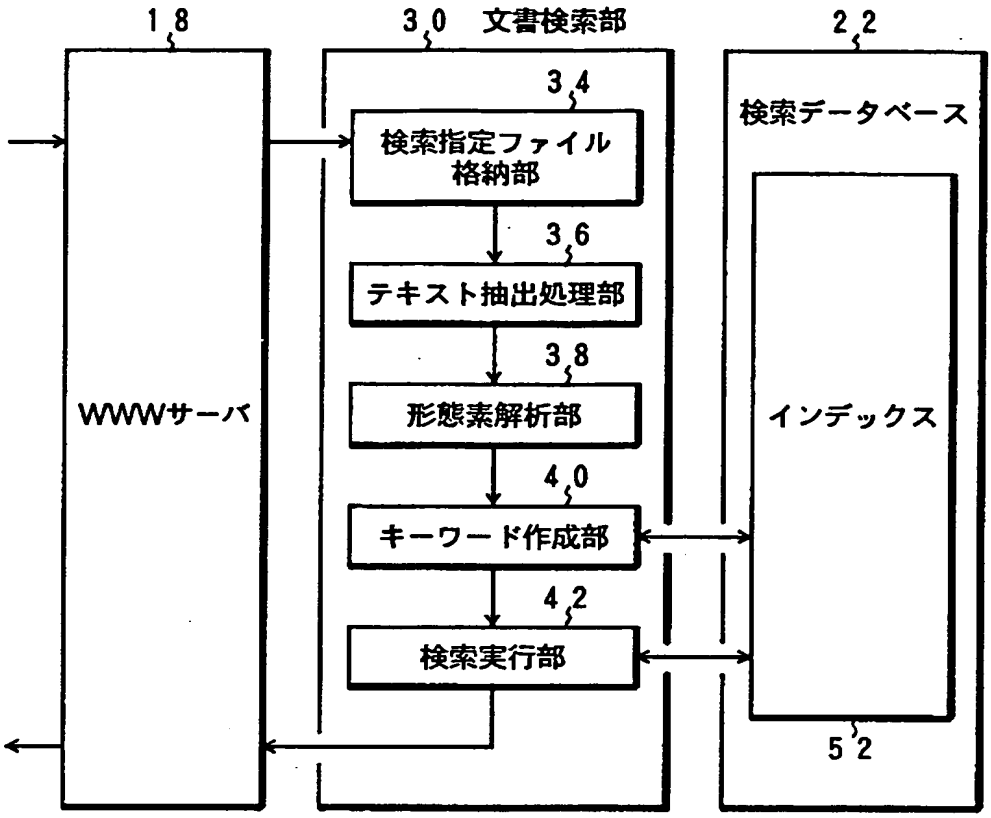
【図３】

本発明の機能構成のブロック図



【図４】

本発明による文書検索部のブロック図



【図５】

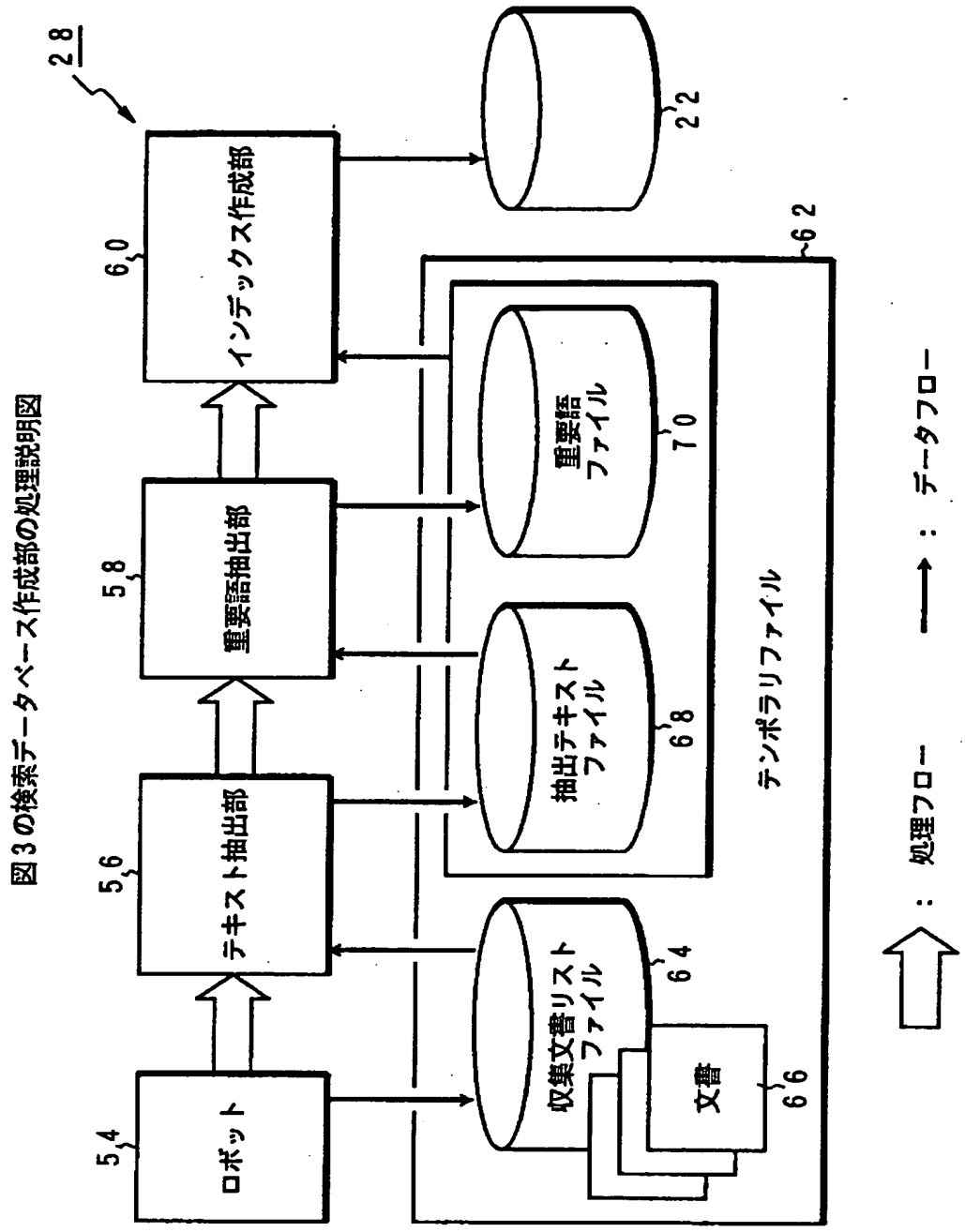
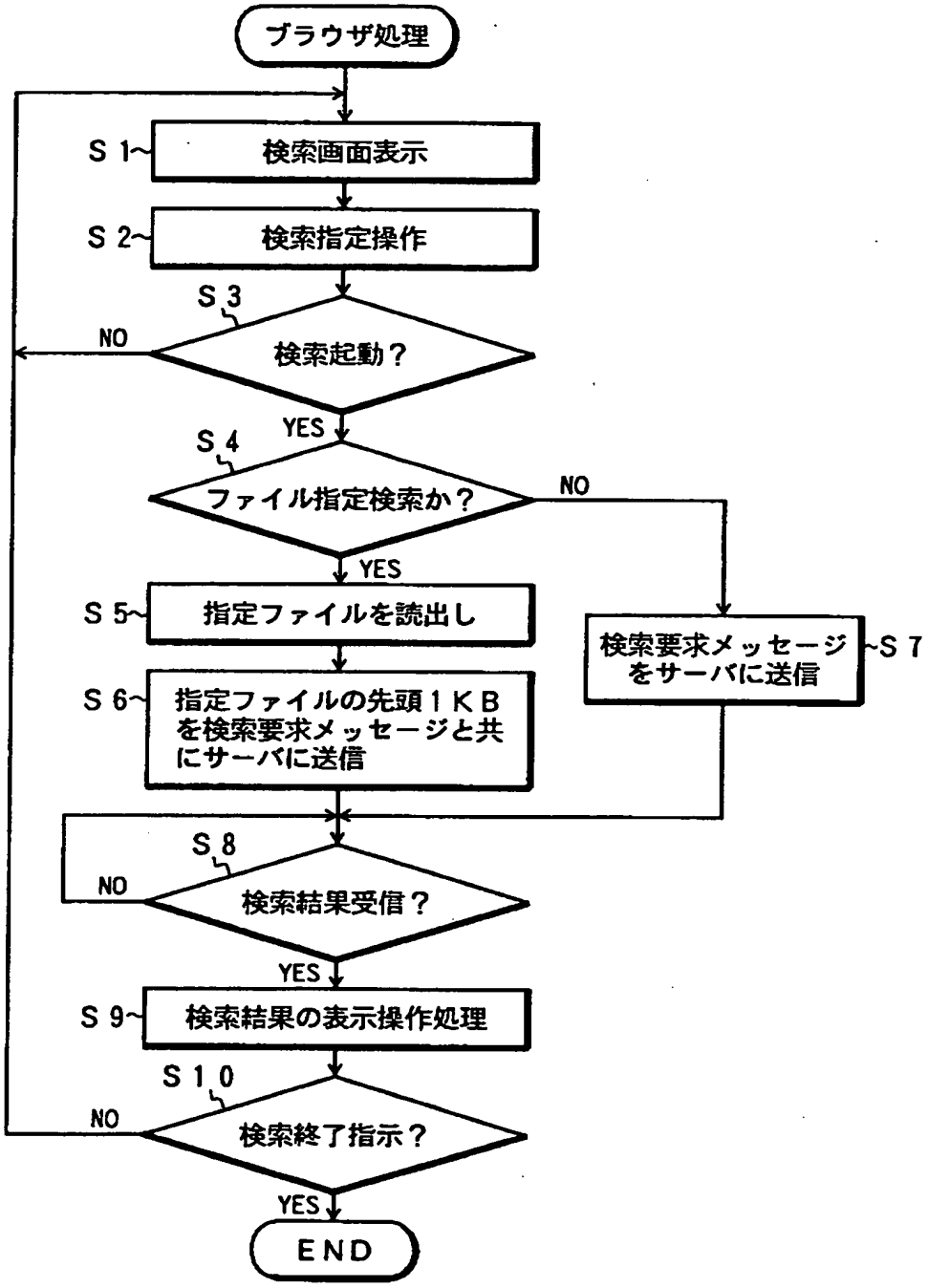


図３の検索データベース作成部の処理説明図

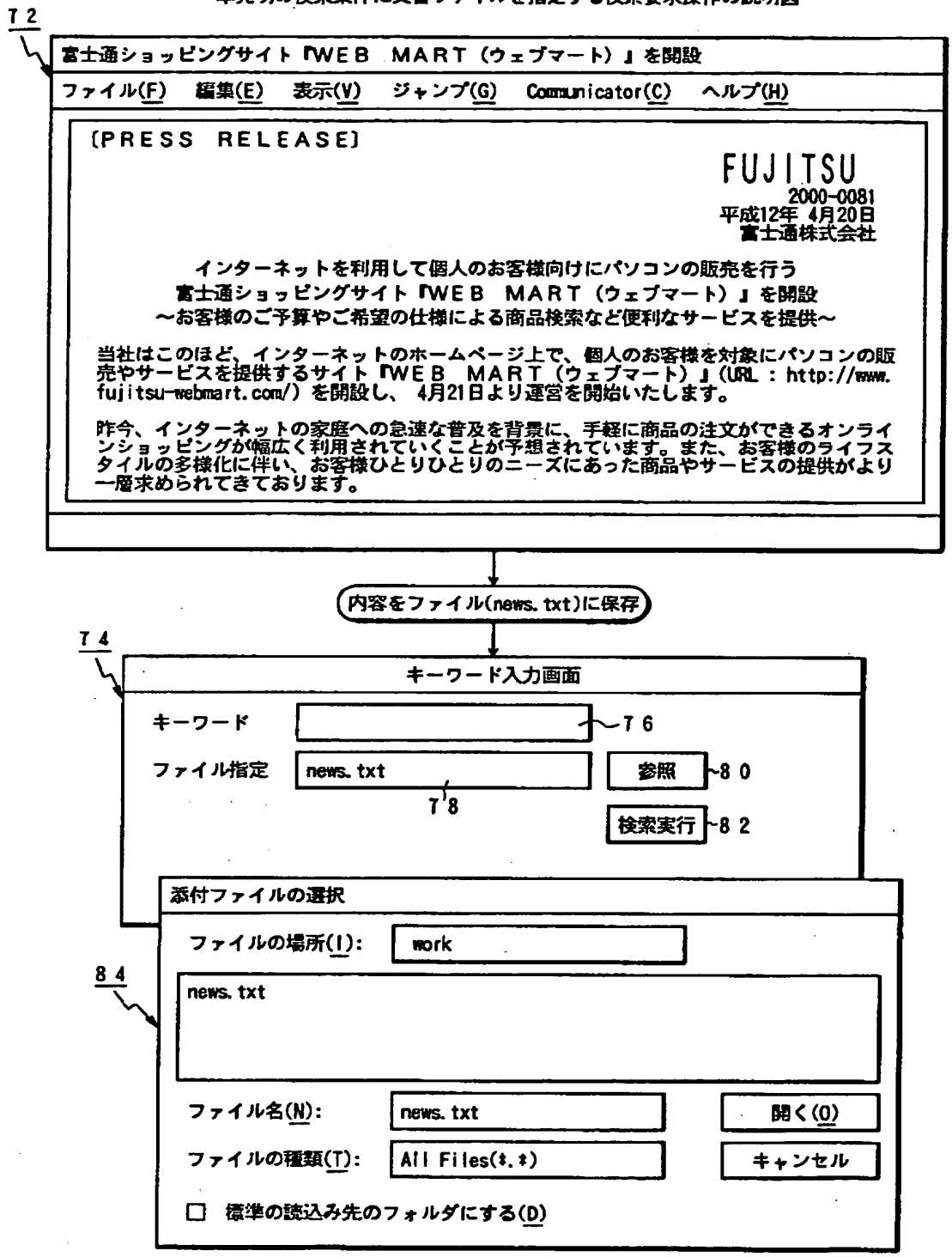
【図６】

図３のブラウザ処理のフローチャート



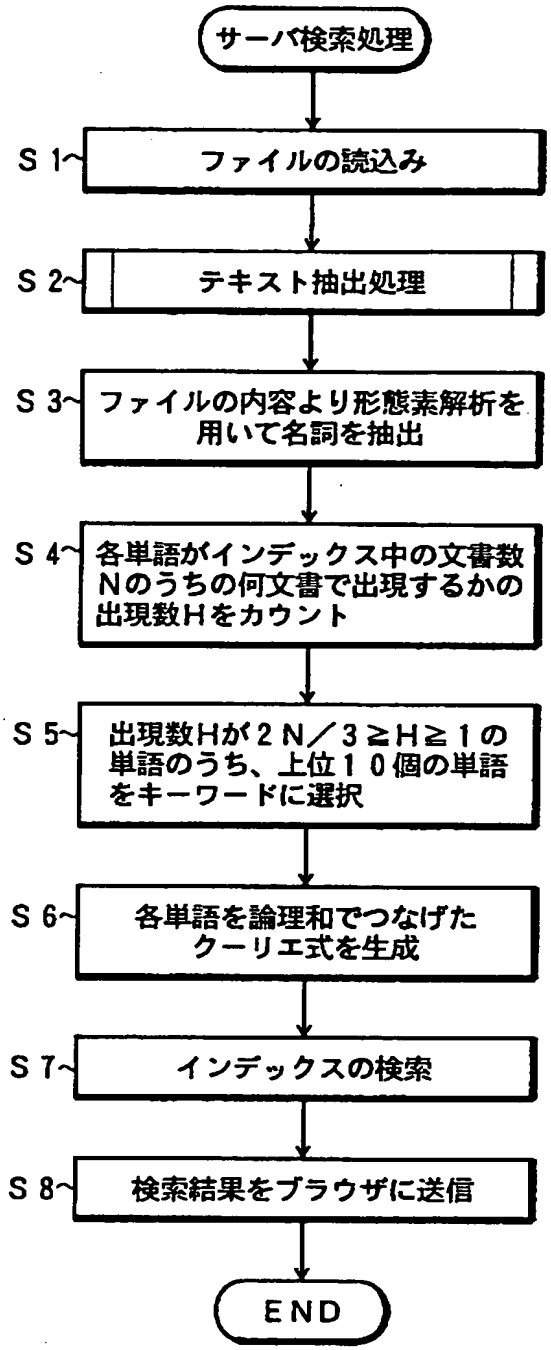
【図７】

本発明の検索条件に文書ファイルを指定する検索要求操作の説明図



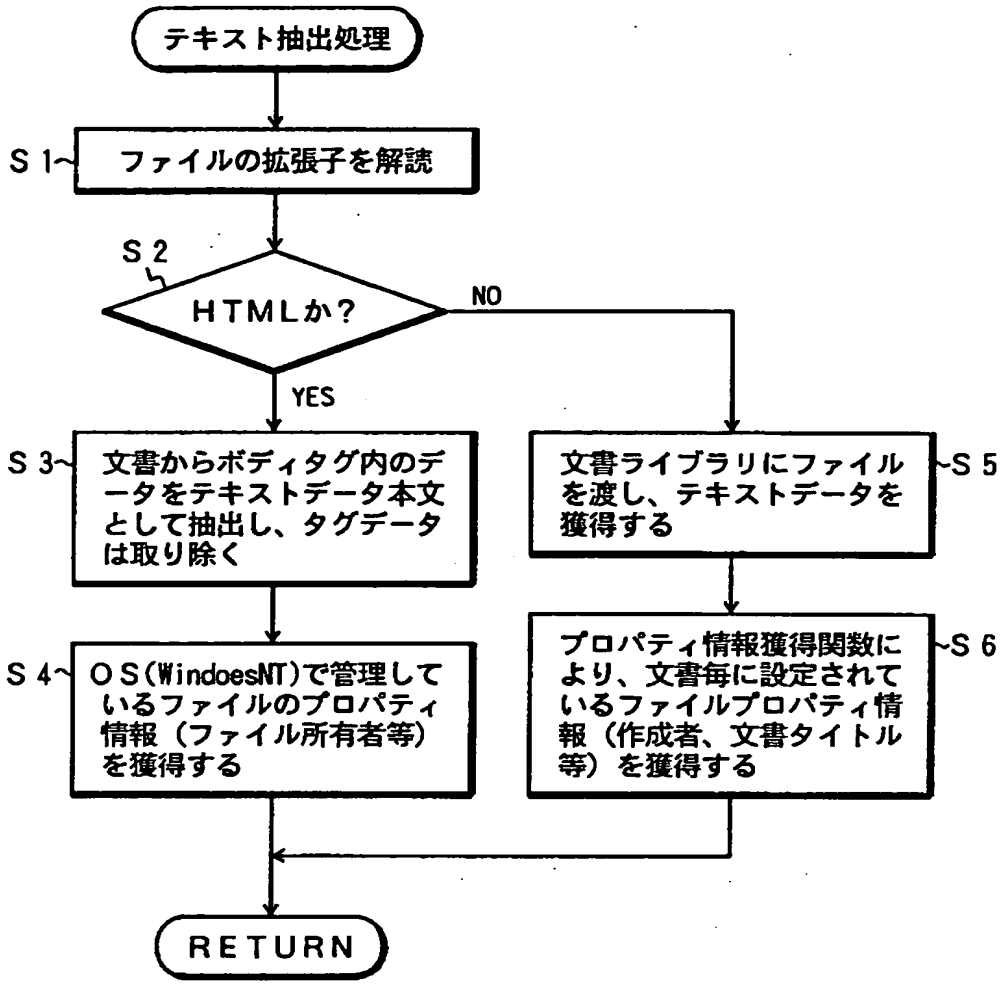
【図８】

本発明のサーバ検索処理のフローチャート



【図９】

図８のテキスト抽出処理のフローチャート



【図１０】

図８の処理によりＨＴＭＬファイルからのテキスト文書を抽出する説明図

(A)

ＨＴＭＬファイル

```
<HTML>
<HEAD>
<ＴＩＴＬＥ>検索データベース</ＴＩＴＬＥ>
</HEAD>
<BODYBGCOLOR="#ＦＦＦＦＦＦ">
<Ｈ４>検索データベース</Ｈ４>
<HR>

<Ｐ>検索に使用するデータベースのことです。<ＢＲ>
検索対象の文書のタイトルや保管場所が記録されています。検索データベースは定期的に
更新され、最新の情報が記録されます。IntelligentSearch は検索データベースにより、
高速検索を実現しています。<ＢＲ>


</BODY>
</HTML>
```

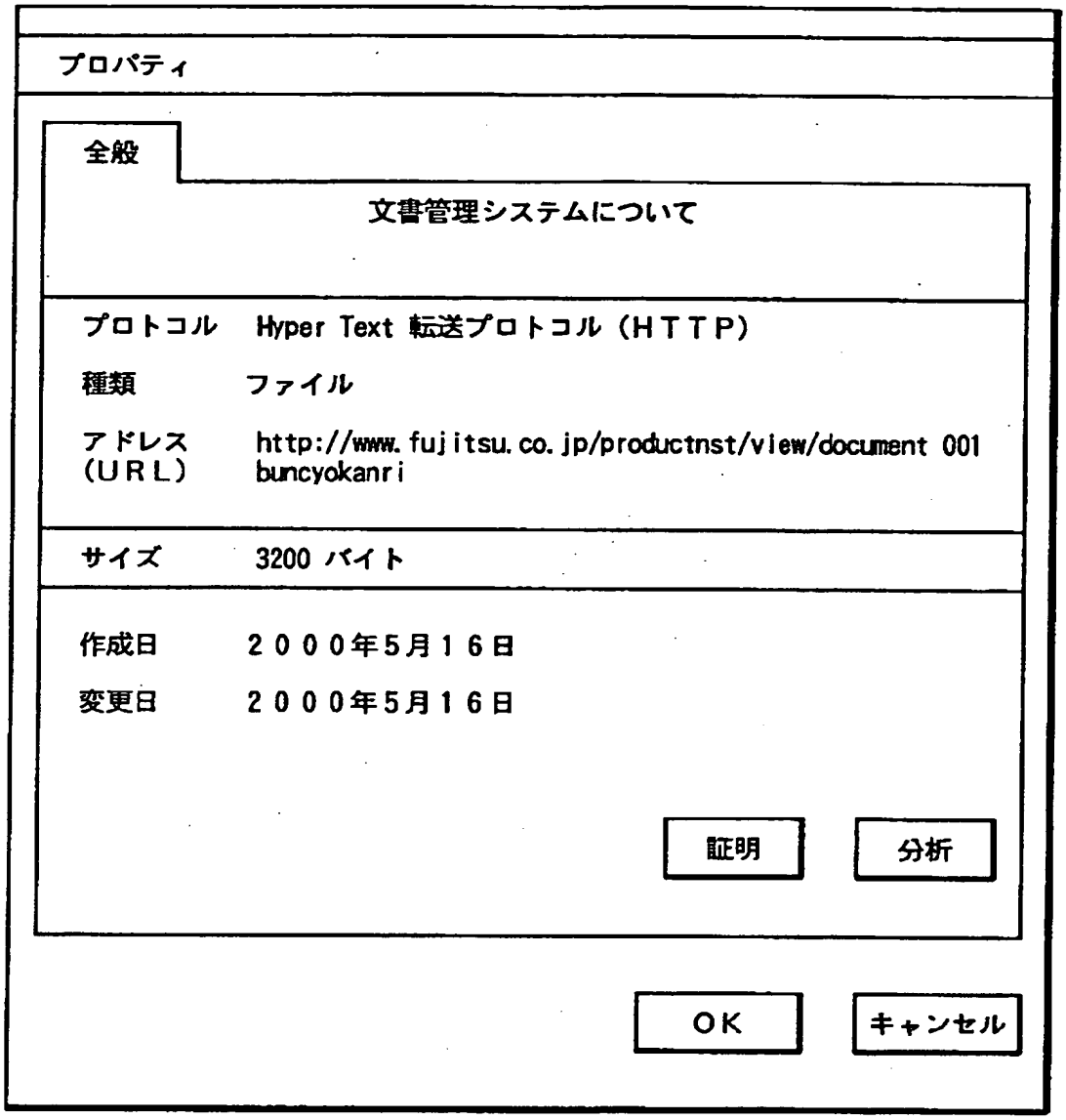
(B)

抽出テキスト文書

検索データベース検索データベース検索に使用するデータベースのことです。検索対象の文書のタイトルや保管場所が記録されています。検索データベースは定期的に更新され、最新の情報が記録されます。IntelligentSearch は検索データベースにより、高速検索を実現しています。

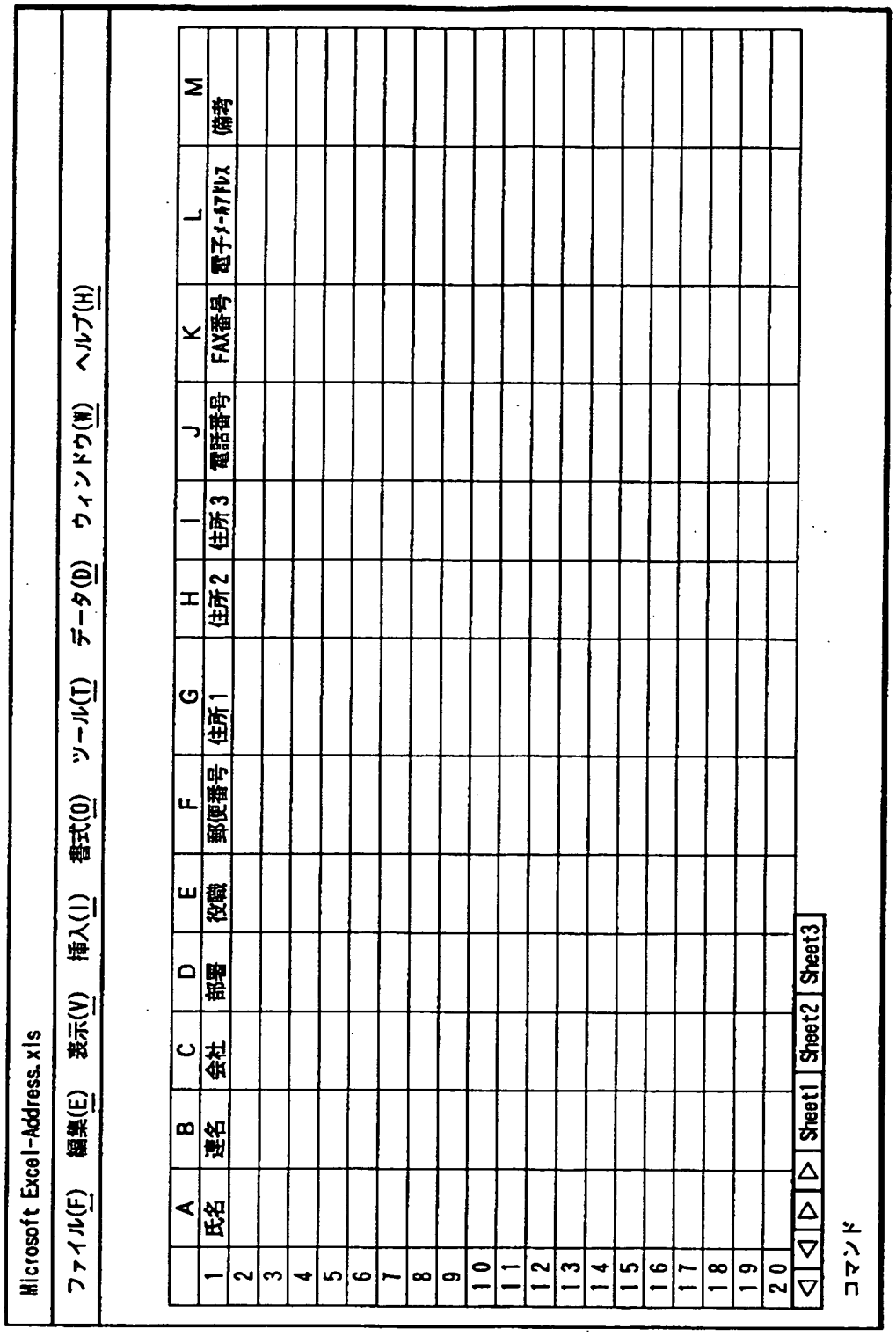
【図１１】

本発明の検索に使用するＨＴＭＬファイルに設けたプロパティ情報の説明図



【図１２】

図８の処理によりテキスト抽出対象とするＥｘｃｅｌ文書の説明図



【図１３】

図１２のＥｘｃｅｌ文書から抽出したテキスト文書の説明図

抽出テキスト文書

"Ｓｈｅｅｔ", "氏名", "連名", "会社", "部署", "役職", "郵便番号",
"住所_１", "住所_２", "住所_３", "電話番号", "ＦＡＸ番号",
"電子メール_アドレス", "Ｓｈｅｅｔ１", "Ｓｈｅｅｔ２", "Ｓｈｅｅｔ３"

【書類名】　　要約書

【要約】

【課題】検索データベースに登録されていない文書に類似した文書の検索をを簡単な操作ですばやく行う。

【解決手段】クライアント１２からの検索要求に基づいてサーバ１０で文書情報を検索して応答する装置であって、クライアント１２の検索条件指定部２６で検索条件に文書ファイルを指定した場合に、指定したファイル内容をネットワークを経由して送信する。サーバ１０側に設けた検索マシン２０の文書検索部３０は、検索条件指定部２６から送信されたファイル内容からキーワードを生成し、検索データベース２２のインデックス（検索対象文書２５から抽出した重要単語列）から類似文書を検索する

【選択図】　　図１

出　願　人　履　歴　情　報

識別番号　　〔０００００５２２３〕

１．変更年月日　　１９９６年　３月２６日
［変更理由］　住所変更
住　所　　神奈川県川崎市中原区上小田中４丁目１番１号
氏　名　　富士通株式会社